Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection

Tapani, Karoliina T.

2019-05

Tapani , K T , Vanhatalo , S & Stevenson , N J 2019 , ' Time-Varying EEG Correlations Improve Automated Neonatal Seizure Detection ' , International Journal of Neural Systems , vol. 29 , no. 4 , 1850030 . https://doi.org/10.1142/S0129065718500302

http://hdl.handle.net/10138/311546
https://doi.org/10.1142/S0129065718500302

unspecified
acceptedVersion

International Journal of Neural Systems, Vol. 0, No. 0 (2018) 1–15

TIME-VARYING EEG CORRELATIONS IMPROVE AUTOMATED NEONATAL SEIZURE DETECTION

KAROLIINA T TAPANI
Medical Imaging and Radiation Therapy,
Kymenlaakso Central Hospital,
Kotka, Finland and
Aalto University School of Science, Espoo, Finland.
E-mail: karoliina.tapani@aalto.fi

SAMPSA VANHATALO
Clinicum, Faculty of Medicine,
University of Helsinki,
Helsinki, Finland

NATHAN J STEVENSON
Clinicum, Faculty of Medicine,
University of Helsinki,
Helsinki, Finland and
Brain Modelling Group,
QIMR Berghofer Medical Research Institute,
Brisbane, Australia

The aim of this study was to develop methods for detecting the non-stationary periodic characteristics of neonatal electroencephalographic (EEG) seizures by adapting estimates of the correlation both in the time (spike correlation; SC) and time-frequency domain (time-frequency correlation; TFC). These measures were incorporated into a seizure detection algorithm (SDA) based on a support vector machine to detect periods of seizure and non-seizure. The performance of these non-stationary correlation measures was evaluated using EEG recordings from 79 term neonates annotated by three human experts. The proposed measures were highly discriminative for seizure detection (median AUC_{SC}: 0.933 IQR: 0.821-0.975, median AUC_{TFC}: 0.883 IQR: 0.707-0.931). The resultant SDA applied to multi-channel recordings had a median AUC of 0.988 (IQR: 0.931-0.998) when compared to consensus annotations, outperformed two state-of-the-art SDAs ($p<0.001$) and was non-inferior to the human expert for 73/79 of neonates.

Keywords: electroencephalography; support vector machines; time-frequency distributions; neonatal seizure detection; nonstationary signal processing.

1. Introduction

Seizures are a common emergency in the neonatal intensive care unit (NICU). They have been associated with increased damage to the developing brain and, therefore, need to be reliably detected to guide treatment and determine prognosis.[1,2] Thus far, visual interpretation of long-duration electroencephalographic (EEG) recordings has been the gold standard of seizure detection, as the majority of seizures do not have clear clinical manifestations.[3] As the interpretation of neonatal EEG is time con-

suming and requires specialized expertise, it is not always available on demand. This has driven the development of numerous automated seizure detection algorithms (SDA) since 1992.[4–11] Even though the performance of these algorithms has improved, researchers have yet to determine if their methods reach the benchmark performance of annotation of the EEG by the human expert. A benchmark that is not absolute due to subjectivity between experts.[12]

In automated neonatal seizure detection, signal transformations that emphasize periodicity are crucial to efficiently and reliably distinguish the repetitive characteristics of seizure. Two traditional approaches have been widely applied to represent periodicity: the autocorrelation in the time domain and the Fourier spectrum (FS) in the frequency domain.[4,5,8] These approaches assume stationarity within the EEG signal; assumptions that are not valid as neonatal seizures exhibit non-stationarity with a time-varying period of repetition.[13–16] The effectiveness of the autocorrelation function and the FS to discriminate neonatal seizures from background EEG is, therefore, reduced. To overcome non-stationarity, adaptive segmentation in the time domain, edge linking to align spectral peaks in the time-frequency domain and time-frequency matched filters have been applied.[7,8,16,17]

The aim of this paper is to improve neonatal seizure detection using estimates of autocorrelation that take into account the time-varying periodicity of neonatal seizures. We propose two methods: spike correlation (SC) and time-frequency correlation (TFC). The SC is a maximum cross-correlation with respect to time lag between adaptively extracted segments of the EEG signal, whereas the TFC is a maximum cross-correlation between scale-shifted time-slices of a time-frequency distribution (TFD). The SC is adapted from the methods of Navakatikyan et al. and Deburchgraeve et al.[7,8]

2. Data

The continuous, 18-channel, EEG measurements analyzed in this study were recorded at the NICU of the Children's Hospital, Helsinki University Central Hospital, Finland. Each EEG recording was initially requested based on clinical suspicion of seizure during routine care and the entire, unedited, recording was included in our dataset. The data was recorded from 79 full-term neonates using a NicoletOne vEEG system (sampled at 256 Hz). The median postnatal age of the neonates was 3 (IQR: 2-10) days. Etiologies (in descending order of occurrence) include hypoxic-ischaemic encephalopathy (HIE), stroke, infection, and other brain abnormalities. For recordings, 19 electrodes were placed according to the international 10-20 system with a bipolar montage employed for the analysis: Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fz-Cz and Cz-Pz.[18] On average, the length of each recording was 85 min (range 52-257 min) and the combined length of recordings approximately 112 hours. Data collection was approved by the Institutional Ethics Committee of the Helsinki University Hospital, Finland.

The data was anonymized and then annotated for the presence of seizure with 1s resolution independently by three clinical experts. Each expert was blinded to each other's annotation and the clinical condition of the neonate. The inter-observer agreement (IOA) between experts, measured by Fleiss' kappa, was 0.777 (95% CI: 0.659-0.830). In total, 39 patients had unanimously annotated seizures (342 consensus seizures, in total) and 22 patients were unanimously annotated as having no seizure. Details on the temporal characteristics of seizures annotated by each expert are presented in Table 1.

3. Methods

3.1. *Evaluating periodicity in the EEG*

According to international guidelines, the EEG manifestation of neonatal seizures is defined as: 'clear ictal events characterized by the appearance of sudden, repetitive, evolving stereotyped waveforms with a definite beginning, middle, and end'.[19] A key component of neonatal seizure detection is, therefore, the detection of evolving repetition or periodicity in the EEG.

Periodicity within any signal, $x(t)$, is defined as,

$$x(t) = x(t+T), \tag{1}$$

where t is time and T is the period. This, however, is a strict definition of periodicity and in real world conditions, where signals are quantized, sampled and embedded in noise, impossible to satisfy. A more lenient definition introduces an error term to

Table 1. A summary of the seizures detected by each human expert. IQR denotes interquartile range.

	Expert 1	**Expert 2**	**Expert 3**
Neonates with seizures	46	45	53
Median (IQR) number of seizures	5 (2-12)	6 (2-13.5)	6 (3-10.5)
Seizures in total	402	429	548
Median (IQR) duration of seizures (s)	59.5 (24-138)	79 (35-137)	43 (20-114.5)
Median (IQR) seizure burden (min)	10.2 (4.3-23.7)	15.0 (6.6-30.3)	8.6 (2.1-22.5)

take into account real world conditions, resulting in a definition of 'almost periodic' as,[20]

$$|x(t) - x(t+T)| < \epsilon. \tag{2}$$

In neonatal seizures and many other biological signals, repetition is further complicated by a time-varying, evolving or non-stationary characteristic.[21,22] This variation in time results from many factors including changes in physiological demand and nonlinear effects that underpin the physiological basis of EEG generation. Time-variation in the period can be embedded into the definition of 'almost periodic' as follows:

$$|x(t) - x(t+T(t))| < \epsilon, \tag{3}$$

where $T(t)$ is the time-varying period, which is the reciprocal of the instantaneous frequency.[16] This definition is broad and some constraints must be placed on $T(t)$ for useful interpretation. For instance, the frequency content of $T(t)$ should not overlap the frequency content of $x(t)$ (similar to the application of Bedrosian's theorem to the Hilbert transform of a signal).[23]

Neonatal EEG seizures display two specific types of time-varying periodicity (see Fig. 1 for examples). The first, and most common, are seizures that consist of a sequence of epileptic spikes where the spike morphology does not change significantly over time,

$$x(t) = w(t) \sum_{n=0}^{N-1} \delta(t - t_n), \tag{4}$$

where t_n is a time shift. In this case, $T(t)$ will be discrete and only needs to be defined for the duration of the waveform prototype, $w(t)$, defining the morphology of the epileptic spike. This function, typically, has a form $e^{-t}\sin(t^{-2})$.[21,22]

The second type defines a seizure, where the fundamental waveform is shifted in time and scale:

$$x(t) = \sum_{m=0}^{M-1} a_m \sin\left(2\pi m \int_0^t T(\tau)^{-1} d\tau + \phi_m\right), \tag{5}$$

where τ is also time, ϕ is a phase constant, a_m and m define the harmonic relationship in the signal, $T(\tau)$ is the time-varying period and continuous.[23]

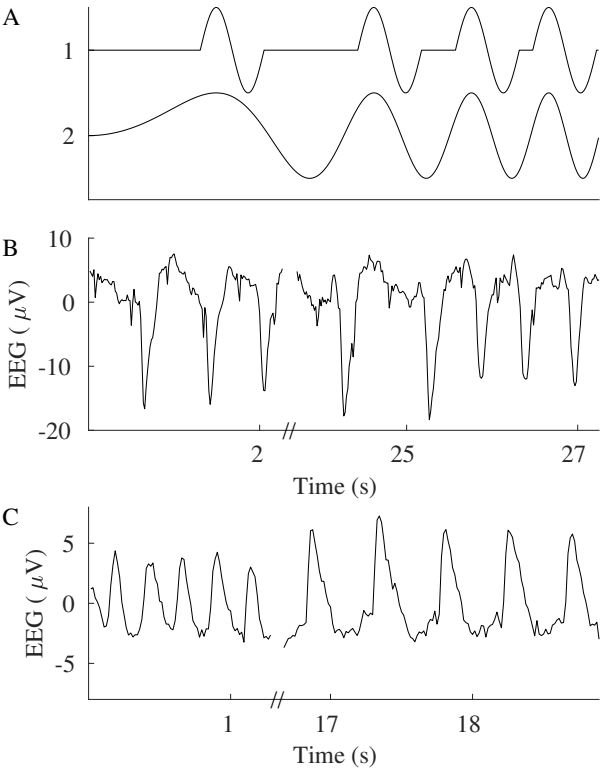


Figure 1. Manifestations of time-varying periodicity in neonatal seizures. A) Time and time-scale shifted version of a simulated waveform prototype. B) Time shifts in the period of a neonatal seizure C) Time shifts in the period and scale shifts in the underlying spike waveform in neonatal seizure. // delineates a break in time.

3.2. *Correlation in time and time-frequency domain*

The estimation of $T(t)$ from a signal is, therefore, fundamental in detecting time-varying periodicity. There are multiple methods for extracting $T(t)$ from a signal,[15,16,24] but the estimation of $T(t)$ does not, however, result in useful detection statistics as $T(t)$ estimated from non-seizure EEG (a filtered $1/f$ process) can overlap with seizure.[25]

A more useful detection statistic is based on the estimate of ϵ in Eq. (3) and it is these measures that we use as the basis of our SDA. A key representation for estimating both T and ϵ from a stationary signal is the autocorrelation function:

$$R_x(\tau) = E[x(t)x(t+\tau)], \tag{6}$$

where $T = \max_\tau\{R_x(\tau)\}$, $\tau > \tau_0$ defines the period and $R_x(T)$ is a surrogate measure of ϵ. This method uses comparisons between the signal and time-shifted versions of itself to determine time instances of maximum similarity which correspond to the signal period. A limitation on τ is required due to correlations within the fundamental waveform and is typically constrained to be greater than τ_0 which is the minimum value of τ where $R_x(\tau) < 0, \tau > 0$.[26] This limit is not considered when evaluating the cross-correlation.

The expectation operator in the definition of $R_x(\tau)$ can be defined as average across time under the assumption of ergodicity and is replaced with a sample mean for discrete time limited signals resulting in a biased estimate of $R_x(\tau)$. $R_x(\tau)$ can also be normalized to the signal variance. The autocorrelation function was the basis of the initial neonatal seizure detection algorithm of Liu et al.[4]

For non-stationary signals, a time-varying form of $R_x(\tau)$ should be applied. This would result in an estimate of $T(t)$ and a surrogate measure of ϵ defined as $R_x(t, T(t))$. In this study, we used adaptive segmentation and time-frequency distributions to estimate $R_x(t, \tau)$.

3.2.1. *Time domain*

Adaptive segmentation has been previously applied in neonatal seizure detection.[7,8] Here, we modify the algorithm of Deburgraeve et al. to provide an implementation for discrete signal epochs.[a] Segmentation is employed to separate single epileptic spikes. The time lag of maximum correlation between successive segments provides the estimate of period and the normalized correlation provides the surrogate measure of ϵ.

The adaptive segmentation is based on the nonlinear energy operator (NLEO),

$$\psi\{x(n)\} = x(n-l)x(n-p) - x(n-q)x(n-s), \tag{7}$$

where $x(n)$ represents discrete EEG signal at sample n, and l, p, q and s are discrete time shifts.[28] We used $l = 1$, $p = 2$, $q = 0$, and $s = 3$ and applied a moving average filter of 7 samples (110 ms). This smoothed NLEO (SNLEO) was then segmented using an adaptive threshold. The optimal threshold was defined as the initial threshold value resulting in largest range of no change to the number of detected bursts (see Fig. 2A3). The search range was 10-90 % quantile of the SNLEO. The correlation of spikes (SC) was used to define a feature for seizure detection as follows:

$$SC_{i,j} = \max_m R_{x_i x_j}(m) \tag{8}$$

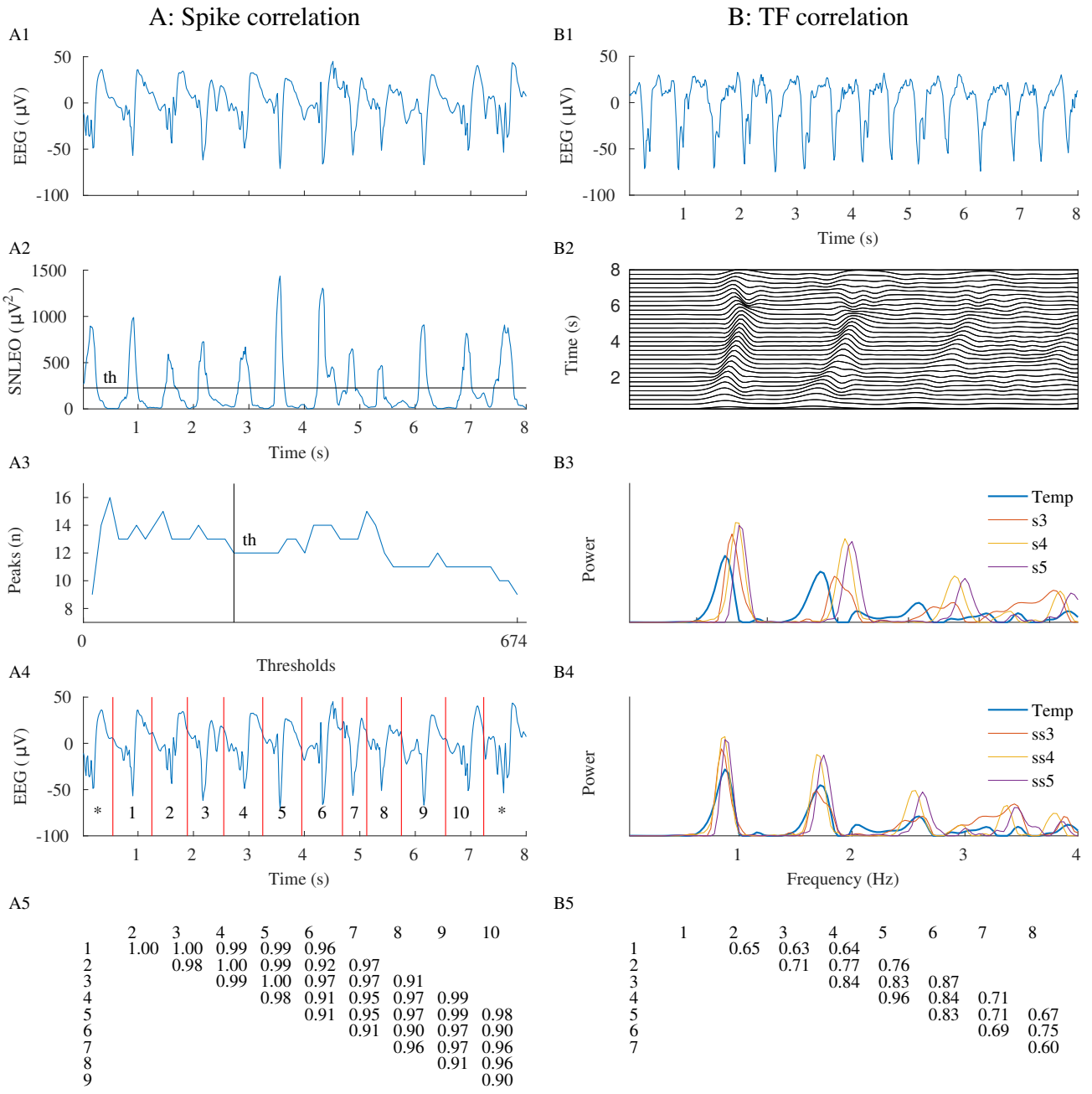
$$R_{x_i x_j}(m; i, j) = \frac{E[x_i(n)x_j(n+m)]}{\sigma_{x_i}\sigma_{x_j}} \tag{9}$$

where x_i and y_j are segments of the EEG signal ($i = [2, ..., S-6]$, $j = [i+1, ...i+5]$), S represents the number of spikes detected in an epoch and σ is the standard deviation of the EEG segment. We use the mean of the SC across EEG segments as a feature for seizure detection. The process of estimating the SC is illustrated in Fig. 2A.

3.2.2. *Time-frequency domain*

Time-frequency distributions (TFD) provide a representation of signal energy over time and frequency, and have been used extensively in neonatal seizure detection.[14,16,17] Cross-correlation is performed between spectra at different times. In this case, the correlation is performed with respect to changes in frequency scale which preserves any harmonic relationships – relationships that were not taken into account in other time-frequency methods.[16] The scale

[a] We have presented elements of the SC algorithm in[27]



	2	3	4	5	6	7	8	9	10
1	1.00	1.00	0.99	0.99	0.96				
2		0.98	1.00	0.99	0.92	0.97			
3			0.99	1.00	0.97	0.97	0.91		
4				0.98	0.91	0.95	0.97	0.99	
5					0.91	0.95	0.97	0.99	0.98
6						0.91	0.90	0.97	0.90
7							0.96	0.97	0.96
8								0.91	0.96
9									0.90

	1	2	3	4	5	6	7	8
1		0.65	0.63	0.64				
2			0.71	0.77	0.76			
3				0.84	0.83	0.87		
4					0.96	0.84	0.71	
5						0.83	0.71	0.67
6							0.69	0.75
7								0.60

Figure 2. Methods for estimating ϵ from signals with time-varying periodicity. A) SC is estimated by adaptively segmenting the EEG signal (A1), using the smoothed NLEO (A2) with an adaptive threshold (A3). The resultant segments of EEG (A4) are then correlated with each other over varying time-shifts and the matrix of maximum correlations (A5) is then summarized to derive a feature. In this example, the mean of SC is 0.96, whereas using the stationary correlation results in a value of 0.36. B) TFC is estimated by comparing time slices of a TFD of the pre-whitened EEG epoch (B2). TFD time-slices (B3) are correlated with scale-shifted versions of future time-slices (B4) to form a matrix of maximum correlations (B5). In this example, the median of TFC is 0.73, whereas using frequency shifts rather than scale shifts results in a value of 0.38.

of maximum correlation at each time of the TFD can be used to estimate the instantaneous frequency, which is related to the time-varying period $T(t)$, and the maximum correlation value is a surrogate measure of ϵ.

The most common TFD, the Wigner-Ville Dis-

tribution (WVD), is defined as,

$$W_z(t,f) = \int P_z(t,\tau)e^{-j2\pi f\tau}d\tau, \tag{10}$$

$$P_z(t,\tau) = z(t+\frac{\tau}{2})z^*(t-\frac{\tau}{2}), \tag{11}$$

where $z(t)$ is an analytic signal and the superscript * denotes the complex conjugate. The WVD is usually convolved (**) with a 2D filter $\gamma(t,f)$ resulting in a smoothed WVD:

$$\rho_z(t,f) = W_z(t,f)^{**}_{tf}\gamma(t,f). \tag{12}$$

In this study, we applied a 2D Hamming window (2.4s in duration and a bandwidth of 1.6Hz). The WVD was then sub-sampled via summation over rectangular regions (32 in time and 4 in frequency) to decrease computation time. We use a smoothed WVD as it has been shown to have good performance when representing neonatal EEG.[29]

The TFD was estimated from the pre-whitened EEG. The whitening filter has a power law response similar to the NLEO at the lower EEG frequency bands.[30] Pre-whitening the spectrum enhances the differences between non-seizure and seizure signals, especially in the high frequency part of the spectrum.[16] This filtering process was implemented by deconvolving,

$$h(k) = \begin{cases} 1, & k = 0 \\ (\frac{\alpha}{2}+k-1)(\frac{h_{k-1}}{k}), & k = [2,...,N] \end{cases} \tag{13}$$

from the signal, where N is length of the signal and $\alpha = 2H+1$ ($H = 2-D_f$ and D_f is estimated using Higuchi's estimate of the fractal dimension).[31,32]

We computed the time-frequency correlation (TFC) as

$$TFC_{i,j} = \max_{\alpha} R_\rho(\beta;i,j) \tag{14}$$

$$R_\rho(\beta;i,j) = \frac{E[\rho(i,k)\rho(j,\frac{k}{\beta})]}{\sigma_{\rho_i}\sigma_{\rho_j}} \tag{15}$$

where $i = [1,...,N-4]$, $j = [i+1,...i+3]$ and N represents the number of time-slices. β is limited to an equivalent frequency shift of ± 1.25 Hz.[15] To generate the scaled time-slices $\rho(j,\frac{k}{\beta})$, we applied interpolation with Hermite splines. We use the median of the TFC across time slices as a feature for seizure detection. The process of estimating the TFC is illustrated in Fig. 2B.

3.3. *Overview of the SDA*

Neonatal seizures are difficult to detect and no single feature has shown sufficient discrimination between seizure and non-seizure.[16,33] We, therefore, selected 19 additional features, *a priori*, to supplement the SC and TFC measures and combined them to form a detection statistic. Features related to the mean SC include the standard deviation of the SC, the mean, standard deviation and skewness of the SNLEO and the number, duration, inter-spike interval, and the mean maximum SNLEO value of detected spikes. We included a measure of SNLEO regularity calculated as the standard deviation of all non-overlapping 2s epochs within 32s of SNLEO output. Features related to the median TFC include relative spectral power, power, measurements of the total harmonic distortion (power in the fundamental, power in the first three harmonics and the maximum value of the power spectral density). We also include a time-varying estimate of the total harmonic distortion defined as,

$$\frac{\sum_{n=0}^{N-1}\max(\rho(n,m))}{\sum_{n=0}^{N-1}\sum_{m=0}^{M-1}\rho(n,m)} \tag{16}$$

where n is discrete time, m is discrete frequency and ρ is the discrete smoothed WVD. This measure is conceptually similar to the feature used outlined in.[16] Finally, we include a measure of the mean EEG amplitude envelope. Short descriptions of each feature are presented in Appendix A Table A.1.

The subsequent SDA is as follows: each channel of neonatal EEG is high-pass filtered with an optimized cutoff frequency (0.5 Hz or 1 Hz, details in Appendix A Table A.2) and a 50 Hz notch filter. The EEG is re-sampled from 256 Hz to 64 Hz and segmented into epochs of 32s in duration, with a 28s overlap (4s time shift). We have shown that this longer epoch slightly improves the ability of features to discriminate between seizure and nonseizure.[27] Features are then estimated from each epoch and combined using a support vector machine (SVM). The sequence of SVM outputs is calculated from the full EEG recording and each channel is post-processed with a moving average filter of 3 samples (12 s) in length. Multi-channel outputs are then transformed into a single decision value using a maximum operator, which is further processed with a median filter of 3 samples (12 s). The duration of

moving average and median filters was selected *a priori* to minimize the influence of a single outlying data point and maximise time resolution. The final decision is computed by applying a threshold to the post-processed SVM output and applying a collar (extending the detection forwards in time). The collar duration was applied to cater for overlap in the epoch segmentation and optimised for each algorithm (the collar duration that maximised SDA performance was selected with a search range of 1-32s). The optimal collar for the proposed SDA was 23s.

3.4. *Training, testing and performance evaluation*

Training and testing of the proposed SDA were performed within a leave-one-patient-out cross validation. In this case, EEG data from 78 patients were included in training set and 1 patient was left out for testing. This process was repeated until every patient had been tested, resulting in 79 trained SVMs. The training set contained in average 543 consensus seizure and 5882 consensus non-seizure epochs per patient (in total 20373 seizure epochs and 458796 non-seizure epochs for each training iteration). The SVM was implemented using the fitcsvm function in Matlab (version R2017a) with hyperparameter optimization (C and σ; Bayesian optimisation within an internal 5-fold cross validation). Hyperparameters and support vectors were, therefore, selected using only training data. For the proposed SDA, the median (IQR) of the box constraints were $10^{2.325}$ ($10^{1.323-2.777}$) and kernel scales were $10^{1.116}$ ($10^{0.828-1.243}$). Features were normalised at each training iteration to z-scores.

In order to determine if the individual features provide a useful level of discrimination between seizure and nonseizure, their ability as an SDA was evaluated using the area under the receiver operator characteristics (AUC) summarised across patients.[34] We apply post-processing stages to each feature for fair comparison with subsequent multi-feature, SVM based detectors.

The performance of SDAs were assessed by comparing the output to the consensus annotations of human experts using temporal and event based measures: AUC, and seizure detection rate (SDR) at false detections or false positives per hour (FD/h) of 0 and 1.[5,35] These values were summarized across patients with consensus seizures (n=39). To include all patients into analysis, AUCs were also calculated using a concatenated annotation (all annotations linked together to form a single annotation approximately 112 h in length). The characteristics of missed seizures and false seizure detections were analysed.[34]

The algorithm was also compared to the annotations of the human experts using measures of IOA to determine its sufficiency. We define a sufficient SDA as an algorithm that generates an annotation of the EEG that is non-inferior to the human expert taking into account the subjectivity of human annotation. It is evaluated using bootstrap estimates of the differences in Fleiss' kappa (κ) between an 'all human' annotation (three human experts) and a composite 'human/SDA' annotation (two human experts and the SDA).[36] Fleiss' kappa statistic is an estimate of percentage agreement between the annotation of more than 2 observers that takes into account the possibility of chance agreement due to the distribution of the data. It is normalised whereby 0 is poor agreement and 1 is perfect agreement. In order to measure SDA sufficiency, we first perform a random sampling (with replacement) of annotations where a patient was considered as a sample. Fleiss kappa was calculated on the concatenated annotations of the 'human/SDA' sample and subtracted from the 'all human' kappa value resulting in $\Delta\kappa$. One thousand random samplings were performed to generate a distribution of $\Delta\kappa$ from which the confidence interval could be estimated. The process was performed on all possible combinations of human experts and SDA (three combinations). If the 95% confidence interval of this distribution spanned zero, it was assumed that the SDA annotation was no different from the annotation of the human expert.

The performance of the SDA was also compared to the algorithm of Deburchgraeve et al. ($\mathrm{SDA_{DB}}$) and to the algorithm of Temko et al. ($\mathrm{SDA_{T}}$).[8,10] We do not use the authors implementation of these algorithms as they are not publicly available. Rather, we use our own implementations based on the original publications trained on our own datasets where possible. As the algorithm of $\mathrm{SDA_{DB}}$ in its original form is not amenable to training, we also implemented a modified version of this algorithm ($\mathrm{SDA_{mDB}}$) where pertinent aspects of $\mathrm{SDA_{DB}}$ are extracted as standalone features that can be trained to form a continuous decision output. In order to ensure valid com-

parisons we apply the same post-processing stages including optimized collars within each SDA implementation. Details on these implementations can be found in Appendix A.

For SDA_{DB}, the AUC was approximated using Hermite splines. SDAs were compared using the Wilcoxon signed rank test on the consensus AUC (across seizure patients; n = 39) and the 95% CI of AUC differences estimated on the concatenated recording using a bootstrap (n=79). P-values less than 0.05 and CIs that did not span zero were deemed to result from significant differences between SDAs.

4. Results

4.1. *Single feature performance*

Single feature AUCs are presented in Table 2 for the 10 best performing features from all SDAs. The mean of SC, the standard deviation of SC, as well as the median of TFC and the time-frequency total harmonic distortion are within the best performing features out of 85 trialled features.

Table 2. The top ten best performing seizure detection algorithms based on a single EEG feature. AUC is the area under the receiver operator characteristic with consensus annotation as the gold standard. Superscripts denote the multi-feature algorithm, in which the feature is included: a - the proposed algorithm, b - the SDA_{mDB} and c - the SDA_T. SC is spike correlation, TFC is time-freqyency correlation, BP is band power, and THD_{TF} is the time-frequency, total harmonic distortion.

Feature	**Mean**	**Median(IQR)**
Mean of $SC^{a,b}$	0.877	0.933 (0.821-0.975)
Spectral Power $(\log)^{a,c}$	0.862	0.899 (0.775-0.958)
Median $TFC^{a,b}$	0.823	0.883 (0.707-0.931)
Amplitude Envelope^{a}	0.778	0.749 (0.650-0.933)
Std of SC^{a}	0.774	0.767 (0.679-0.898)
Relative BP (6-8Hz)^{c}	0.770	0.777 (0.680-0.859)
Shannon Entropy^{c}	0.767	0.751 (0.622-0.919)
THD^{a}_{TF}	0.765	0.760 (0.617-0.905)
Relative BP (7-9Hz)^{c}	0.760	0.776 (0.646-0.849)
Wavelet Power 4-8 Hz^{b}	0.759	0.722 (0.619-0.912)

4.2. *Algorithm performance on full recording*

The AUC of the proposed algorithm was significantly higher than SDA_{DB}, SDA_{mDB} and SDA_T (see Table 3, Fig. 3). SDA_{DB} had a significantly lower AUC than SDA_{mDB} and SDA_T ($p< 0.001$), while SDA_T significantly outperformed SDA_{mDB} ($p< 0.01$). A one-point AUC estimate of each SDA resulted in median AUCs of 0.921 (IQR: 0.709-0.986), 0.772 (IQR: 0.610-0.941) and 0.833 (IQR: 0.609-0.961), for the proposed SDA, SDA_{mDB} and SDA_T, respectively. These values are directly comparable to the AUC values for SDA_{DB} in Table 3. At a threshold that maximised the IOA between the SDA and the human experts on the concatenated annotations, the proposed SDA had a median sensitivity of 0.761 (IQR: 0.399-0.961; n=39), a median specificity of 0.992 (IQR: 0.960-1.000; n=79), a median positive predictive value of 0.720 (IQR: 0.283-0.911; n=39) and a median negative predictive value of 0.972 (IQR: 0.852-1.000; n=79).

The annotations of SDAs were significantly different from that of the human expert (incorporating SDA annotations as an additional expert significantly reduces κ, Table 4). The proposed SDA, however, achieved the benchmark of IOA between human experts if 6 of the worst performing neonates were removed. In contrast, SDA_T, SDA_{mDB}, and SDA_{DB} reached the benchmark when 60, 65, and 74 patients were removed, respectively (Table 4).

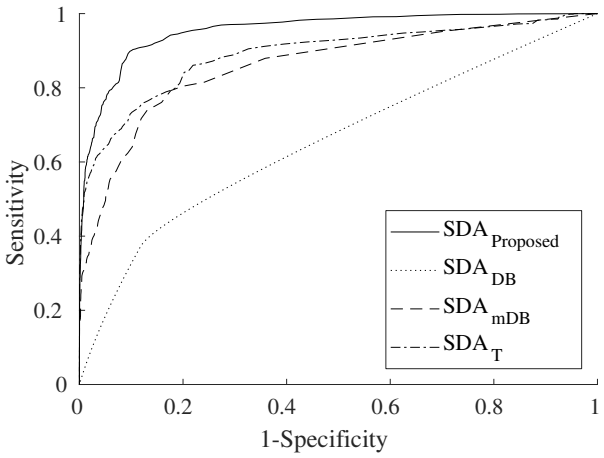


Figure 3. SDA performance against human consensus (n=39, neonates with seizure): sensitivity as a function of 1-specificity, averaged over seizure patients.

The proposed SDA does not perfectly align with the consensus annotation of the human experts. There were 342 consensus seizures detected in 39 neonates. The proposed SDA detected 228 of these seizures in 36 neonates with 30 false detections oc-

Table 3. SDA performance compared to human consensus. Average AUCs and SDRs are computed on consensus seizure patients only (n=39) for consensus seizures of at least 10 seconds. For SDA_{DB}, median SDR is 60.0 % (IQR: 0.6-95.3, n=39) and median FD/h 0 (IQR: 0-1.772, n=79). The p-values are from a Wilcoxon signed rank test between an SDA and the proposed SDA (n=39), the Δ AUC 95% CI is the difference between the AUC_{cc} of an SDA and the proposed SDA estimated using a bootstrap (n=79). AUC_{cc} denotes AUC computed on concatenated recordings.

Algorithm	**Proposed SDA**	**SDA_{DB}**	**SDA_{mDB}**	**SDA_{T}**
Median AUC (IQR)	0.988 (0.931-0.998)	0.683 (0.500-0.818)	0.943 (0.851-0.988)	0.961 (0.869-0.990)
Mean AUC	0.957	0.660	0.886	0.923
AUC_{cc}	0.955	0.767	0.862	0.901
Mean SDR (FD/h$\sim$0)	79.8 %	-	65.9 %	72.2 %
Mean SDR (FD/h$\sim$1)	86.6 %	-	73.5 %	78.4 %
p-values	-	$<$0.001	$<$0.001	$<$0.001
ΔAUC 95% CI	-	0.128-0.283	0.042-0.149	0.028-0.081

curring in 11 neonates. The detected seizures were apparent on a median of 5 channels (IQR: 2-9), with false detections apparent on a median of 1 channel (IQR: 1-2). The distribution of seizure detections and false seizure detections across channels is shown in Fig. 4B. Examples of seizures that were correctly detected, seizures that were missed and false seizure detections are shown in Fig. 5. Missed seizures tended to have a low amplitude and false detections tended to contain distinct delta activity and appear on a lower number of specific channels. Detection accuracy was also dependent on seizure duration (see Fig. 4A).

5. Discussion

Our results suggest that measures of non-stationary correlation provide superior discrimination between seizure and non-seizure in the neonatal EEG, compared to over 50 other EEG features. This implies that there is still room for the development of features for neonatal seizure detection. Incorporating these features into an SDA resulted in superior detection performance in contrast to two state-of-the-art methods on a database of approximately 1 h recordings from 79 term neonates with mixed etiologies. The output annotation of the proposed SDA achieved the benchmark of IOA between human experts with a subset of 73 neonates from the cohort.

The proposed measures of non-stationary correlation provide an estimate of the error term (ϵ) from a definition of time-varying periodicity. This parameter of periodicity is more useful for discriminating between seizure and non-seizure than high precision estimates of the time-varying period. The proposed measures do not rely on internally set thresholds and output a continuous, bounded variable. As such, these measures are highly suited to modern classifiers which can be trained on data sets of labeled EEG. Out of these two measures, the SC outperforms TFC for seizure detection when assessed individually. This suggests a predominance of seizure types with discrete changes in period, resulting from limited spatial dynamics in the underlying seizure. At the biophysical level, it is commonly assumed that waveform shapes relate to different spatio-temporal configurations between the cortical source and the recording electrode (or their paired derivation). For instance, the transition from sharp to smooth waveforms relates to the spread from a focal to wider cortical area, or to shifting of the focus away from the recording electrode. The complexity of translation between cortical event and EEG waveform precludes accurate explanation, however, clinical studies support the idea that a wide range of spatio-temporal seizure configurations should be targeted by an ideal SDA.[37,38]

The proposed SDA significantly outperformed our implementations of two state-of-the-art algorithms when applied to our data set.[8,10] Performance of SDA_{DB} was likely compromised by the use of many fixed thresholds within the algorithm.[8] Reformulating the algorithm so that these thresholds could be trained on our data (SDA_{mDB}) significantly improved performance. Our present findings are, nevertheless, compatible with an idea that SDA_{mDB}, contains too few features, while SDA_{T} includes many features that do not add value for seizure detection. Although we are confident that the implementations

Table 4. Differences of automated annotations to human annotation evaluated using inter-observer agreement. For each SDA, one of the three annotators was replaced by the SDA annotation, resulting in three combinations. The table presents $\Delta\kappa$ values with the 95% confidence intervals in brackets. The average agreement between all iterations of two human annotators and an algorithm (Mean κ) are also presented. Cohort size defines the number of patients with which the IOA benchmark is achieved, after discarding the patients with the lowest agreement for each SDA.

Left Out Expert	**Proposed SDA**	**SDA$_{\mathbf{DB}}$**	**SDA$_{\mathbf{mDB}}$**	**SDA$_{\mathbf{T}}$**
1	0.121 (0.054-0.205)	0.290 (0.218-0.367)	0.202 (0.124-0.300)	0.172 (0.101-0.254)
2	0.095 (0.022-0.190)	0.263 (0.190-0.343)	0.182 (0.099-0.287)	0.143 (0.068-0.238)
3	0.131 (0.065-0.211)	0.294 (0.221-0.372)	0.211 (0.139-0.301)	0.176 (0.107-0.256)
Mean κ	0.646	0.268	0.554	0.590
Cohort size	73 (92%)	5 (6%)	14 (18%)	19 (24%)

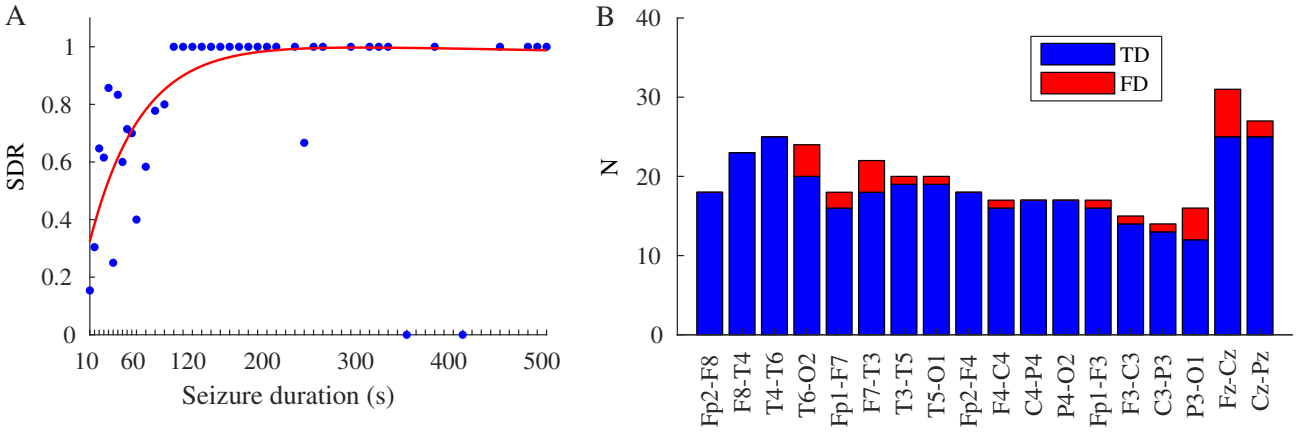


Figure 4. The effect of temporal and spatial distributions on the performance of the proposed seizure detection algorithm. A) The distribution of seizure detection rate (SDR) with seizure duration. B) The distribution of detections across EEG channels. All detections were assessed at the optimal patient independent threshold, N denotes the number of patients, FD is false seizure detections and TD is true seizure detections.

of these algorithms are accurate to the limits of what can be extracted from the published reports, it is important to note that prior SDAs have not been made openly available to allow for direct comparison at the code level. It should also be noted that these algorithms have been improved with continued research.[11,39]

The overall performance metrics reported for the proposed algorithm are similar to those reported in other recent studies.[39,40] Our AUCs may be, nevertheless, overestimated, as we evaluated SDA performance on the consensus annotation only. Traditional performance measures (AUC, SDR and FD/h) are feasible when comparing an SDA output to a gold standard. In the case of neonatal seizure detection, however, the gold standard is subjective and other measures are required to determine whether a SDA is sufficient. Improved measures have been developed that deal with variability in the annotations of the human expert,[41] however, these methods of assessment lack the capacity to determine the sufficiency of an SDA. This raises the question, given the subjectivity of the annotation of the human expert, what value of sensitivity/specificity or seizure detection rate/false detection rate is required to determine if an SDA annotation is non-inferior to the human expert (human equivalent). The IOA measures presented in this study constitute an innovative way of determining the performance of an SDA where the ultimate goal is to generate an annotation of the neonatal EEG that is indistinguishable from the human expert. They are conceptually sim-

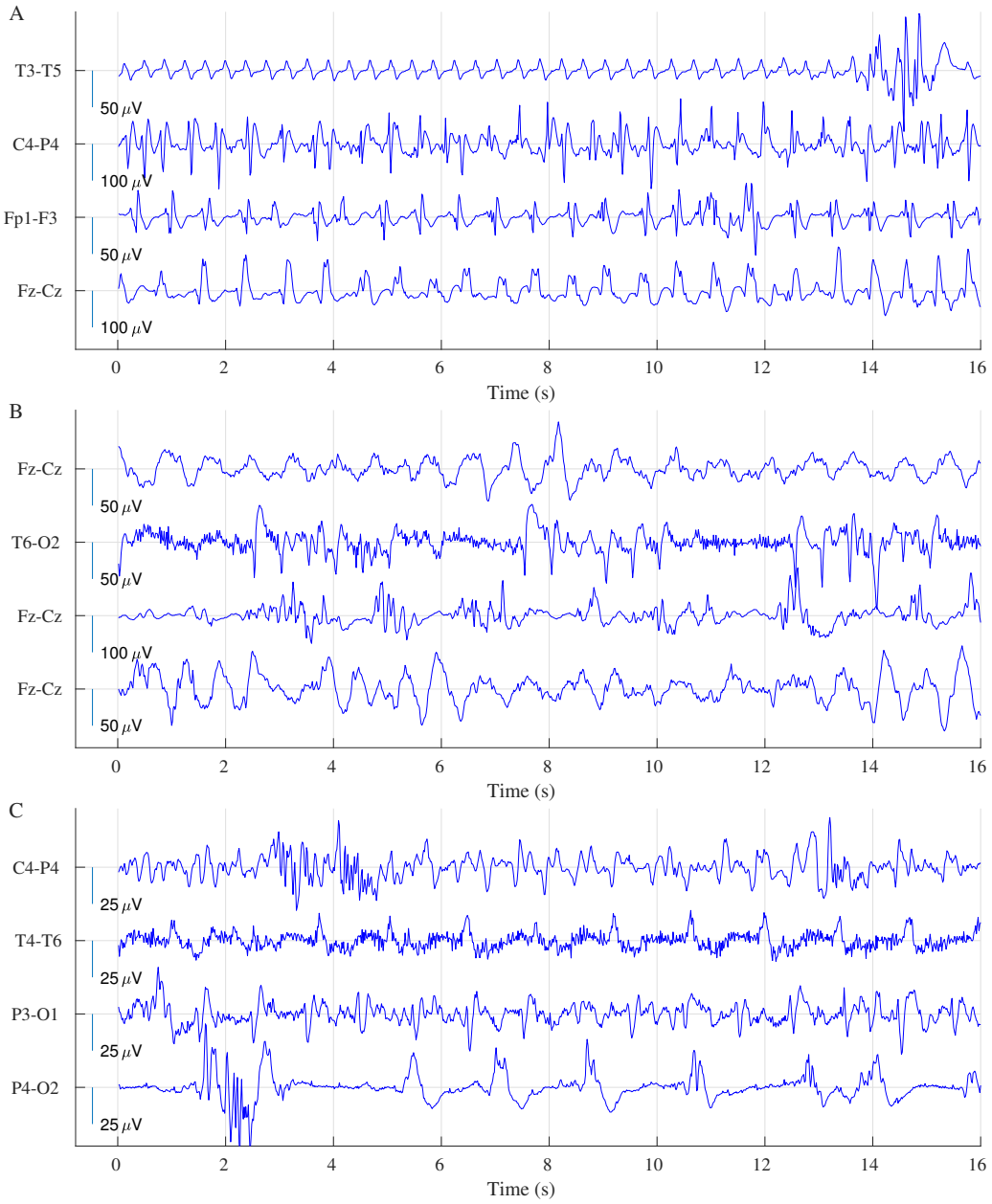


Figure 5. Examples of EEG seizure detection. A) true seizure detections, B) false seizure detections and C) missed seizures. Each trace was from a different neonate and the vertical line at the beginning of each trace denotes the voltage scale. Voltage scales were varied to highlight the morphology of the EEG waveform from different neonates. Note the preponderance of delta waves and Fz-Cz in false detections, and the low amplitude of missed seizures.

ilar to the noninferiority analysis of Scheuer et al.[42] The advantages of our method is that it is based on a single univariate metric rather than multiple bivariate metrics which means that our analysis is not complicated by multiple values and the potential covariance between variables. It is important to also note that while the methods of analysis are valid, a consensus on the level of confidence required, the size of the dataset and the reviewers used has yet to be reached.[43]

The recordings of our study were of short duration and taken at times when there was clinical suspicion of seizures. This results in a significantly higher hourly seizure burden (occurrence over time)

compared to typical long term EEG recordings, the likely common user scenario for future SDA implementations. This is not ideal as there will be limited examples of the many artefacts that can contaminate long duration recordings (potential false detections), no examples of natural changes in the EEG recording environment over the longer term and limited granulation of event based measures (resulting in minimal changes in SDR over a range of false detections from 0 or 1). There may be, nevertheless, some advantages when using short duration recordings. There will be a more amenable balance between seizure and nonseizure data which is useful for training and performance evaluation and there is limited opportunity for long periods of relatively benign EEG patterns to improve measures of specificity (decrease false detections per hour). While the recording durations are relatively low, the number of neonates is high and comparable to published data sets.[7,9] This suggests that our data set provides a valuable and representative sample as inter-subject variability is higher than intra-subject variability in neonatal seizures. The clear advantage of our data set is that it employs the annotations of multiple experts which permits our analysis of sufficiency based on IOA. The open evaluation of data sets, annotations and code remains a problem in the development of neonatal SDAs. We, therefore, make our data, annotations and code publicly available at Zenodo and GitHub.[44–46]

6. Conclusions

We have developed a neonatal SDA based on a set of 21 features combined by a kernel SVM. The feature set contains novel features for indirectly estimating ϵ from a definition of time-varying periodicity. We have also developed a novel method of assessing SDA sufficiency based on measures of IOA. The proposed SDA outperforms our implementation of leading methods with an AUC across all concatenated EEG recordings of 0.955. We have, furthermore, demonstrated two important findings 1) there is still potential for the development of features for neonatal SDAs with an emphasis on time-varying methods and 2) assessments against the benchmark of human subjectivity is the only way to determine the sufficiency of an SDA. Potential improvements to the algorithm include revisiting features used in adult SDAs, incorporating features that can discriminate between seizure and slow repetitive activity such as respiration artefact or delta waves and the use of optimal montages or channel adaptive post-processing stages.[47–49] Prospective validation of the proposed algorithm is also required to determine the generalisability of the proposed SDA and the dataset it was trained on as well as its clinical utility.

Acknowledgments

The authors would like to thank Dr Leena Lauronen and Mr Jarmo Mäki for annotating the data. This study was supported by the Finnish Cultural Foundation, EU Marie Skłodowska-Curie Action (H2020-MCSA-IF-656131), Academy of Finland (#276523 and #288220), Helsinki University Hospital, Sigrid Juselius Foundation, the National Health and Medical Research Council Australia (#1144936) and Rebecca L Cooper Foundation.

Conflict of interest

The authors declare no conflicts of interest.

Appendix A

Features

The Table A.1 shortly describes all features used in the proposed SDA. These features consist of the SNLEO features (utilizing adaptive segmentation with the smoothed non-linear energy operator), frequency spectrum features (computed from the Welch's power estimate and short-time Fourier transform), the median of time-frequency correlations from the TFD and amplitude envelope.

Algorithm implementations

In this paper, we implemented two versions of the method of Deburchgraeve et al.:[8] 1) the original algorithm according to the first publication, $\mathrm{SDA_{DB}}$ and 2) a modified and discretized version, $\mathrm{SDA_{mDB}}$. This modified $\mathrm{SDA_{DB}}$ algorithm is a reformulated version of the original algorithm that can be trained on our data set.[8] We reformulated the algorithm so as to be able to extract features which were necessary to adopt a version of the $\mathrm{SDA_{DB}}$, that would be comparable to common epoch-by-epoch based algorithms. Extracted features that corresponded to

Table A.1. Short descriptions of features in the proposed algorithm, first the features for which optimal high-pass filter cutoff frequency was 1 Hz and second 0.5Hz.

Features	**Short description**
Cutoff 1 Hz	
Skewness	Skewness of the SNLEO ouput
Regularity	Std of skewness of 2s sub-windows of the SNLEO
Spike number	Number of spikes in an epoch
Spike width	Median width of spikes
Spike gap	Median inter-spike interval in
Mean of SC	Mean correlation between spikes
Std of SC	Std of correlation between spikes
Mean SNLEO	Mean of the SNLEO
Std SNLEO	Std of the SNLEO
Spikiness	Mean of spike peaks over background
Median of TFC	Median correlation between scaled time-slices of TFD
Amplitude envelope	Mean of the amplitude envelope
Spectral power (log)	Natural logarithm of the spectral power between 0.5-30 Hz
Cutoff 0.5 Hz	
THD1	Power in the first three harmonics divided by sum of the PSD
THD2	Power in the fundamental divided by the sum of the PSD
THD3	Logarithm of the the power in the fundamental
Relative delta power	0.5-4 Hz
Relative theta power	4-8 Hz
Relative alpha power	8-12 Hz
Relative beta power	12-30 Hz
THD_{TF}	Sum of max TFD of each slice divided by sum of the TFD

important components of the original algorithm are listed in Table A.3.

Table A.2. Filters and epoch lengths. SDA_{mDB} and SDA_{T} also applied a notch filter at 50 Hz.

	$\textbf{SDA}_{\textbf{DB}}$	$\textbf{SDA}_{\textbf{mDB}}$	$\textbf{SDA}_{\textbf{T}}$
Sampling frequency (Hz)	256	64	32
High-pass cutoff (Hz)	0.3	1	0.5
Low-pass cutoff (Hz)	30	32	8
Epoch length (s)	-	32	8
Optimal collar length (s)	28	28	7

We also implemented the algorithm of Temko et al. (SDA_{T}) for comparison on our data; the features of the SDA_{T} can be found in.[10] For the SVM training of these algorithms, the median optimal hyperparameters, box constraint and kernel scale parameters were selected as $10^{2.437}$ (IQR: $10^{1.184-2.979}$; SDA_{mDB}) and $10^{0.481}$ (IQR: $10^{0.257-0.591}$; SDA_{T}).

Bibliography

1. S. Björkman, S. Miller, S. Rose, C. Burke and P. Colditz, Seizures are associated with brain injury severity in a neonatal model of hypoxia–ischemia, *Neuroscience* **166**(1) (2010) 157–167.
2. S. Miller, J. Weiss, A. Barnwell, D. Ferriero, B. Latal-Hajnal, A. Ferrer-Rogers, N. Newton, J. Partridge, D. Glidden, D. Vigneron *et al.*, Seizure-associated brain injury in term newborns with perinatal asphyxia, *Neurology* **58**(4) (2002) 542–548.
3. D. M. Murray, G. B. Boylan, I. Ali, C. A. Ryan, B. P. Murphy and S. Connolly, Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures, *Arch Dis Child-Fetal* **93**(3) (2008) F187–F191.
4. A. Liu, J. Hahn, G. Heldt and R. Coen, Detection of neonatal seizures through computerized EEG analysis, *Electroen Clin Neuro* **82**(1) (1992) 30–37.
5. J. Gotman, D. Flanagan, J. Zhang and B. Rosenblatt, Automatic seizure detection in the newborn: Methods and initial evaluation, *Electroen Clin Neuro* **103**(3) (1997) 356–362.
6. P. Celka and P. Colditz, A computer-aided detection of EEG seizures in infants: A singular-spectrum approach and performance comparison, *IEEE T Bio-Med Eng* **49**(5) (2002) 455–462.
7. M. A. Navakatikyan, P. B. Colditz, C. J. Burke,

Table A.3. Features derived from the original Deburchgraeve algorithm. Asterisks represent features that are also adopted in the proposed algorithm. DWT stands for discrete wavelet transform, where we used bi-orthogonal wavelet with decomposition order 3 .

Feature	**Description**
Spike number*	Number of spikes in an epoch
Spike width*	Median width of spikes
Mean of SC*	Mean correlation between spikes (up to 5 spikes)
Spikiness*	Mean of spike peaks over background
Skewness(autocorr)	Skewness of autocorrelation function
Wavelet Coeff 1-2 Hz	DWT Coefficient corresponding 1-2 Hz
Wavelet Coeff 2-4 Hz	DWT Coefficient corresponding to 2-4 Hz
Wavelet Coeff 4-8 Hz	DWT Coefficient corresponding to 4-8 Hz
Zero-crossings(autocorr)	Difference of intervals between zero-crossings in the autocorrelation [%]

T. E. Inder, J. Richmond and C. E. Williams, Seizure detection algorithm for neonates based on wave-sequence analysis, *Clin Neurophysiol* **117**(6) (2006) 1190–1203.

8. W. Deburchgraeve, P. Cherian, M. De Vos, R. Swarte, J. Blok, G. Visser, P. Govaert and S. Van Huffel, Automated neonatal seizure detection mimicking a human observer reading EEG, *Clin Neurophysiol* **119**(11) (2008) 2447–2454.
9. J. Mitra, J. R. Glover, P. Y. Ktonas, A. T. Kumar, A. Mukherjee, N. B. Karayiannis, J. D. Frost Jr, R. A. Hrachovy and E. M. Mizrahi, A multi-stage system for the automated detection of epileptic seizures in neonatal EEG, *J Clin Neurophysiol* **26**(4) (2009) 218–226.
10. A. Temko, E. Thomas, W. Marnane, G. Lightbody and G. Boylan, EEG-based neonatal seizure detection with support vector machines, *Clin Neurophysiol* **122**(3) (2011) 464–473.
11. A. Ansari, P. Cherian, A. Dereymaeker, V. Matic, K. Jansen, L. De Wispelaere, C. Dielman, J. Vervisch, R. Swarte, P. Govaert *et al.*, Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor, *Clin Neurophysiol* **127**(9) (2016) 3014–3024.
12. N. J. Stevenson, R. R. Clancy, S. Vanhatalo, I. Rosén, J. M. Rennie and G. B. Boylan, Interobserver agreement for neonatal seizure detection using multichannel EEG, *Ann Clin Transl Neurol* **2**(11) (2015) 1002–1011.
13. P. Celka, B. Boashash and P. Colditz, Preprocessing and time-frequency analysis of newborn EEG seizures, *IEEE Eng Med Biol* **20**(5) (2001) 30–39.
14. B. Boashah and M. Mesbah, A time-frequency approach for newborn seizure detection, *IEEE Eng Med Biol* **20**(5) (2001) 54–64.
15. L. Rankine, N. Stevenson, M. Mesbah and B. Boashash, A nonstationary model of newborn EEG, *IEEE T Bio-Med Eng* **54**(1) (2007) 19–28.
16. N. J. Stevenson, J. M. OToole, L. J. Rankine, G. B. Boylan and B. Boashash, A nonparametric feature for neonatal EEG seizure detection based on a representation of pseudo-periodicity, *Med Eng Phys* **34**(4) (2012) 437–446.
17. M. S. Khlif, M. Mesbah, B. Boashash and P. Colditz, Multichannel-based newborn EEG seizure detection using time-frequency matched filter, *IEEE EMBS*, 2007, pp. 1265–1268.
18. N. Stevenson, L. Lauronen and S. Vanhatalo, The effect of reducing EEG electrode number on the visual interpretation of the human expert for neonatal seizure detection, *Clin Neurophysiol* **129**(1) (2018) 265–270.
19. R. R. Clancy and A. Legido, The exact ictal and interictal duration of electroencephalographic neonatal seizures, *Epilepsia* **28**(5) (1987) 537–541.
20. C. Corduneanu, *Almost periodic functions* (Chelsea Pub Co, 1989).
21. N. J. Stevenson, M. Mesbah, G. B. Boylan, P. B. Colditz and B. Boashash, A nonlinear model of newborn EEG with nonstationary inputs, *Ann Biomed Eng* **38**(9) (2010) 3010–3021.
22. S. B. Nagaraj, N. J. Stevenson, W. P. Marnane, G. B. Boylan and G. Lightbody, Neonatal seizure detection using atomic decomposition with a novel dictionary, *IEEE T Bio-Med Eng* **61**(11) (2014) 2724–2732.
23. B. Boashash, *Time-frequency signal analysis and processing: a comprehensive reference* (Academic Press, 2015).
24. B. Boashash, Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals, *P IEEE* **80**(4) (1992) 520–538.
25. N. Stevenson, I. Korotchikova, A. Temko, G. Lightbody, W. Marnane and G. Boylan, An automated system for grading EEG abnormality in term neonates with hypoxic-ischaemic encephalopathy, *Ann Biomed Eng* **41**(4) (2013) 775–785.
26. J. Theiler, Spurious dimension from correlation algorithms applied to limited time-series data, *Phys Rev A* **34**(3) (1986) p. 2427.

27. K. Tapani, S. Vanhatalo and N. Stevenson, Incorporating spike correlations into an SVM-based neonatal seizure detector, *EMBEC*, **65**2017, pp. 322–325.
28. J. O'Toole, A. Temko and N. Stevenson, Assessing instantaneous energy in the EEG: A non-negative, frequency-weighted energy operator, *IEEE EMBC*, 2014, pp. 3288–3291.
29. B. Boashash and S. Ouelha, Automatic signal abnormality detection using time-frequency features and machine learning: A newborn EEG seizure case study, *Knowl-Based Syst* **106** (2016) 38–50.
30. J. M. OToole, G. B. Boylan, R. O. Lloyd, R. M. Goulding, S. Vanhatalo and N. J. Stevenson, Detecting bursts in the EEG of very and extremely premature infants using a multi-feature approach, *Med Eng Phys* (2017).
31. T. Higuchi, Approach to an irregular time series on the basis of the fractal theory, *Physica D* **31**(2) (1988) 277–283.
32. N. J. Kasdin, Discrete simulation of colored noise and stochastic processes and $1/f^{\alpha}$ power law noise generation, *P IEEE* **83**(5) (1995) 802–827.
33. B. Greene, S. Faul, W. Marnane, G. Lightbody, I. Korotchikova and G. Boylan, A comparison of quantitative EEG features for neonatal seizure detection, *Clin Neurophysiol* **119**(6) (2008) 1248–1261.
34. A. Temko, E. Thomas, W. Marnane, G. Lightbody and G. Boylan, Performance assessment for EEG-based neonatal seizure detectors, *Clin Neurophysiol* **122**(3) (2011) 474–482.
35. S. B. Wilson, M. L. Scheuer, C. Plummer, B. Young and S. Pacia, Seizure detection: Correlation of human experts, *Clin Neurophysiol* **114**(11) (2003) 2156–2164.
36. J. L. Fleiss, Measuring nominal scale agreement among many raters., *Psychol Bull* **76**(5) (1971) 378–382.
37. A. Bye and D. Flanagan, Spatial and temporal characteristics of neonatal seizures, *Epilepsia* **36**(10) (1995) 1009–1016.
38. M. D. Bourez-Swart, L. van Rooij, C. Rizzo, L. S. de Vries, M. C. Toet, T. A. Gebbink, A. G. Ezendam and A. C. van Huffelen, Detection of subclinical electroencephalographic seizure patterns with multichannel amplitude-integrated EEG in full-term neonates, *Clin Neurophysiol* **120**(11) (2009) 1916–1922.
39. A. Temko, G. Boylan, W. Marnane and G. Lightbody, Robust neonatal EEG seizure detection through adaptive background modeling, *Int J Neural Syst* **23**(04) (2013) p. 1350018.
40. E. Thomas, A. Temko, G. Lightbody, W. Marnane and G. Boylan, Gaussian mixture models for classification of neonatal seizures using EEG, *Physiol Meas* **31**(7) (2010) 1047–1064.
41. A. H. Ansari, P. Cherian, A. Caicedo, K. Jansen, A. Dereymaeker, L. De Wispelaere, C. Dielman, J. Vervisch, P. Govaert, M. De Vos *et al.*, Weighted performance metrics for automatic neonatal seizure detection using multi-scored EEG data, *IEEE J Biomed Health* (2017) p. 10 pages.
42. M. L. Scheuer, A. Bagic and S. B. Wilson, Spike detection: Inter-reader agreement and a statistical Turing test on a large data set, *Clin Neurophysiol* **128**(1) (2017) 243–250.
43. M. B. Westover, J. J. Halford and M. T. Bianchi, What it should mean for an algorithm to pass a statistical Turing test for detection of epileptiform discharges, *Clin Neurophysiol* **128**(7) (2017) 1406–1407.
44. Matlab code: `https://github.com/ktapani/Neonatal_Seizure_Detection`.
45. EEG dataset, DOI:10.5281/zenodo.1280684 `https://zenodo.org/record/1280684#.Wxh3QkiFNaQ`.
46. Trained SVMs, DOI:10.5281/zenodo.1281146 `https://zenodo.org/record/1281146#.WxjW7nVubCI`.
47. H. Adeli, S. Ghosh-Dastidar and N. Dadmehr, A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy, *IEEE T Bio-med Eng* **54**(2) (2007) 205–211.
48. J. Li, W. Zhou, S. Yuan, Y. Zhang, C. Li and Q. Wu, An improved sparse representation over learned dictionary method for seizure detection, *Int J Neural Syst* **26**(01) (2016) p. 1550035.
49. S. Yuan, W. Zhou, Q. Wu and Y. Zhang, Epileptic seizure detection with log-Euclidean Gaussian kernel-based sparse representation, *Int J Neural Syst* **26**(03) (2016) p. 1650011.